Using temporal public transport demand profiles to reveal urban spatial patterns

Alonso Espinosa Mireles de Villafranca*[1], Zhiren Huang[2], and Charalampos Sipetas[3]

[1]Postdoctoral researcher, Department of Civil Engineering, Tampere University, Finland
[2]Postdoctoral researcher, Department of Computer Science, Aalto University, Finland
[1,2]Postdoctoral researcher, Department of Built Environment, Aalto University, Finland

Short summary

We present a versatile method, inspired by computational neuroscience, for reconstructing smooth demand profiles from sparse timestamp data for public transport boardings. We show how areas can be clustered based on the similarity of their temporal demand profiles to reveal urban spatial patterns. We use the Helsinki metropolitan region to showcase the method using data on boarding events from the TravelSense data from HSL (the Helsinki region transport authority) collected through their mobile ticketing app. Our results show the show the dependence of travel demand on available public transit and modes and supply volume. Furthermore, the clusters align with extremely well with the types of urban areas in the region. Due to the high supply and even frequency of transit options, the differences in demand profiles are due to mode availability and land-use features rather than frequency patterns.
Keywords: Demand Patterns, Mobile Data, Multi-modal Transport, Public Transport

1 Introduction

Mobility patterns reflect the inherent complexity of human behaviour. Spatially, their characterisation ranges from large scale coverage (Gonzalez et al., 2008) to the individual bus stop level.
Aggregated travel behaviour gives rise to emergent patterns that are tied to space and the built environment. Features of these patterns can reveal complex spatial relationships such as polycentric urban activity (Roth et al., 2011) and basins of attraction for journeys (Cats et al., 2015). Moreover, the behavioural aspects of journeys are spatio-temporal. Trip purpose depends on the time of day, and the origin and destination depend on the purpose for travelling.
Previously, models for inferring activities and trip purposes were mainly based on location and time gaps (Devillaine et al., 2012; Lee & Hickman, 2014) avoided the wider built environment. While valuable in transport planning contexts, their validation requires large amounts of data. However, the increasing availability of large and rich datasets enables improving models that attempt to capture complex behavioural drivers for travel (Cuauhtemoc Anda & Fourie, 2017), as well as spatio-temporal features of trips in greater detail.
Urban form and the built environment are important in determining travel behaviour (Ewing & Cervero, 2010; Nasri & Zhang, 2012; Hong et al., 2014). As a central urban feature, public transport (PT) underpins travel choices (Miranda-Moreno et al., 2011; Mouratidis et al., 2019) and influences future urban development, especially with increasing implementation of transit-oriented development (Thomas & Bertolini, 2020). Therefore PT, the built environment, and travel patterns are fundamentally interdependent.
Data-based methods, such as clustering techniques, have already proven useful in shedding light on these interdependencies. For example, in understanding how passengers use PT networks, and the networks' functional characteristics. Luo et al. (2017) use construct OD matrices for transit networks, by partitioning stops into clusters on proximity and passenger flows. While Yap et al. (2019) identify transfer hubs using density based clustering taking into account flows of transferring passengers between stops using smartcard data,. In terms of the broader relationship between travel patterns and the urban environment, Cats et al. (2015) identify urban centres based on passenger flows, and in Cats & Ferranti (2022), the authors study their attractiveness and classify urban centres based on activity profiles.

In this paper we present a versatile method for reconstructing localised demand profiles obtained from timestamps of boarding events. Our methodology also covers how to cluster geographical areas based on these profiles to reveal spatial distribution of demand patterns. Due to its generality, as well as being able to reconstruct demand profiles from boarding data, the method can also be used to obtain supply profiles.
To showcase the method, we use the TravelSense dataset (Huang et al., 2022) from the Helsinki metropolitan region to show how PT demand is tied into the urban environment. We further exemplify the method's versatility by comparing demand to the supply of PT services.

2 Methodology

Demand profiles

Our method — inspired by computational neuroscience methods (Houghton, 2015) — takes a series of events (with associated timestamps and obtains a smooth frequency function via a kernel density estimate (KDE) (Wand & Jones, 1994). In our case, the events of interest are passenger boardings of PT vehicles. An advantage over using histograms, is that KDE is not sensitive to the choice of the bin edges, which can introduce errors, especially when comparing distributions.
The KDE demand profile $\hat{f}$ is obtained by summing copies of a *kernel* function (K_h) centred over each timestamp t_i,

$$\hat{f}_h(t) = \frac{1}{n}\sum_{i=1}^{n} K_h(t - t_i), \tag{1}$$

where the kernel is non-negative and scaled by the bandwidth h, which scales the 'width' of the kernel. We use a Gaussian kernel where h corresponds to the standard deviation

$$K_h(t) = \frac{1}{h\sqrt{2\pi}} \exp\left(-\frac{t^2}{2h^2}\right). \tag{2}$$

Supply profiles

Similarly to the demand profiles, we can obtain supply profiles from GTFS data. We can aggregate stops arbitrarily and use the scheduled times of arrival of vehicles as the timestamps to obtain KDE curves as above. However, due to the different transport modes, a weighted KDE has to be used. Each arrival event is weighted by the mean capacity of the vehicle to reflect the PT supply throughout the day. We use capacities reported in HSL report HSL (2021) (Table 1).

Table 1: Average capacities of vehicles in Helsinki by mode HSL (2021)

Mode	Capacity [passengers]
Tram	180
Metro	700
Bus	96
Train	600
Ferry	370

The equivalent expression to 1 with weight factors c_i from Table 1 becomes,

$$\hat{f}_h(t) = \frac{1}{\sum_{i=1}^{n} c_i}\sum_{i=1}^{n} K_h(t - t_i). \tag{3}$$

Clustering

In order to cluster demand profiles together, we use the the L^2 metric (Sutherland, 2009), to quantify how their pairwise similarity. Explicitly, if we have KDE profiles for two different lists, $\hat{f}_h$ and $\hat{g}_h$, the distance between them is given by

$$d(\hat{f}_h, \hat{g}_h) = \|\hat{f}_h - \hat{g}_h\|_2 = \sqrt{\int_a^b \left(\hat{f}_h(\tau) - \hat{g}_h(\tau)\right)^2 d\tau}. \tag{4}$$

We use k-medoid clustering (Kaufman & Rousseeuw, 1990) to group grid cells into categories based on their demand profile. In k-medoids the centre of the cluster is designated as one of the points of the cluster itself.
To select the number of clusters, k, we use the silhouette value (Rousseeuw, 1987) which quantifies how well points fit into their assigned clusters. Higher silhouette values indicate better defined clusters.
For each point i in cluster C_I, its silhouette value $s(i)$ is defined as

$$s(i) = \begin{cases} \dfrac{b(i) - a(i)}{\max\{a(i),\, b(i)\}}, & |c_I| > 1, \\ 0, & \text{otherwise}, \end{cases} \tag{5}$$

where $a(i)$ is the mean distance from i to the points in its cluster and $b(i)$ is the mean distance between i and the points in the nearest neighbouring cluster C_J. Explicitly,

$$a(i) = \frac{1}{|C_\mathrm{I}| - 1} \sum_{j \in C_\mathrm{I}, i \neq j} d(i, j), \tag{6}$$

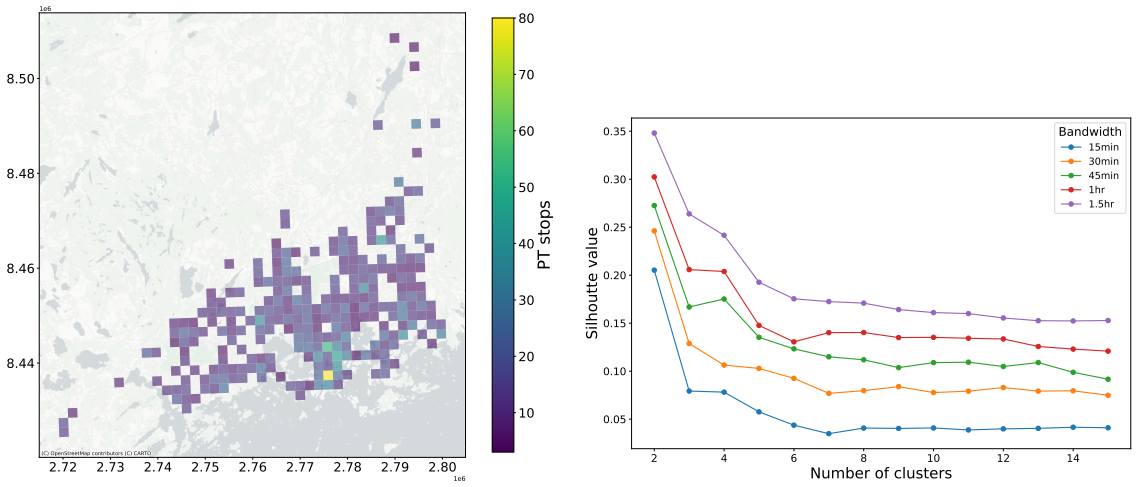
and

$$b(i) = \min_{I \neq J} \frac{1}{|C_\mathrm{J}|} \sum_{j \in C_\mathrm{J}} d(i, j). \tag{7}$$

The silhouette value for a clustering is obtained by averaging the silhouette values of all the points.

Data and preparation

We use the TravelSense dataset for September 2021, collected by HSL (the Helsinki region transport authority) in the Helsinki metropolitan region, to obtain the boardings of PT vehicles at stops (described in Huang et al. (2022)). Temporally, we aggregate all days of the month by consider only time-of-day of events.
We aggregate PT stops over cells of size 1 km^2 and we obtain demand profiles for cells that contain at least 100 timestamps. This left 194 cells across the region, which are visualised in Figure 1a with the number of stops contained in each.



(a) (b)

Figure 1: (a) Cells (of size 1 km^2) that have at least 100 timestamps during September 2021; the colour shows the number of PT stops contained in each cell. (b) Mean silhouette value for different number of clusters and bandwidths (100 k-medoid clusterings were calculated for each curve).

For the supply profiles, we use the GTFS schedule data published by HSL (sourced from HSL (2021), available MobilityData IO (2024)) for the same month as the demand data.

Parameters for clustering

We tested five different bandwidths (h=15min, 30min, 45min, 1hr, 1.5hr) and performed k-medoid clustering for $k = 2$ up to $k = 15$ for each of them. The corresponding silhouette values are shown in figure 1b. For each k-h pair, 100 instances of k-medoid clusterings were obtained and the silhouette value of the clusterings were averaged to account for differences due to the random initial choice of medoids.
For most bandwidth values there is either a step change or a local peak (for $h = 45$ min) in the silhouette value for $k = 4$. We select this number of clusters as a trade-off between high cluster definition (high silhouette score) and having more clusters. The bandwidth value we use in the remainder is $h = 45$ min.

3 Results and discussion

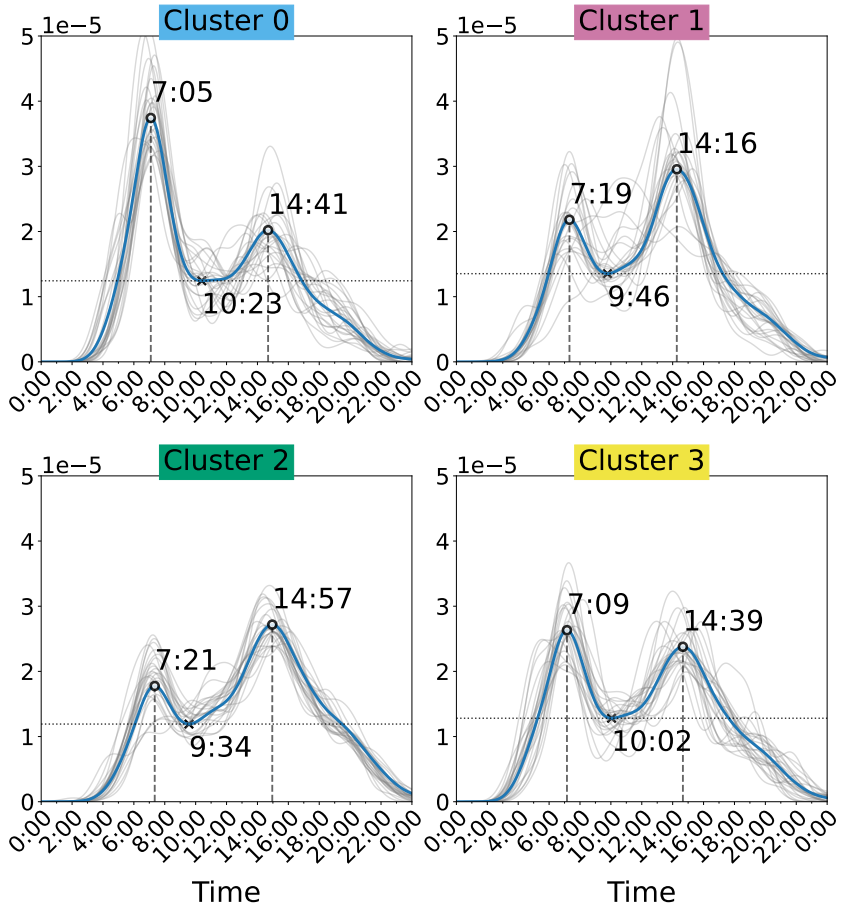
Clustering by demand

We obtain four clusters of grid cells, whose (normalised) demand profiles are shown in Figure 2a. Figure 2b compares the profiles together, scaled according to counts. The spatial distribution of clustered cells are displayed in Figure 3 and Table 2 summarises features of the cluster profiles.

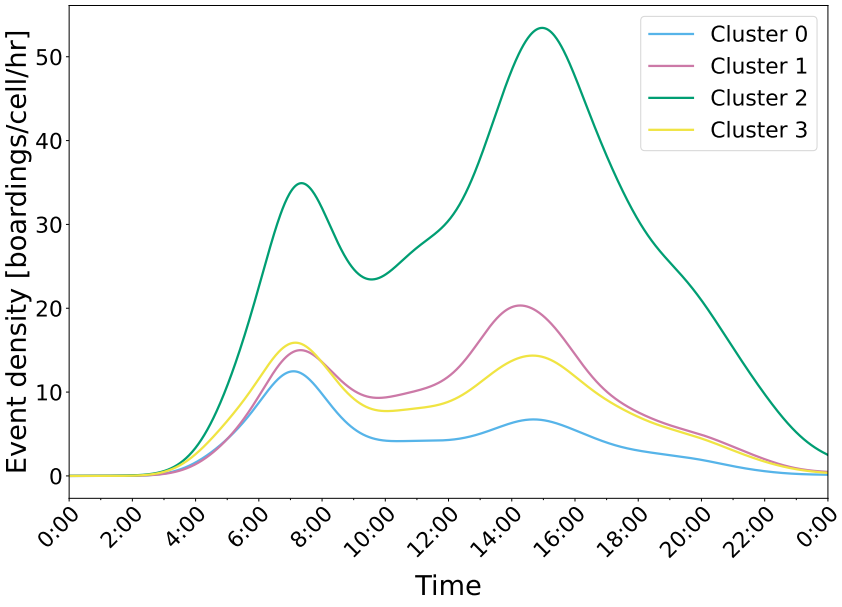
Table 2: Demand profile features for the demand clusters

		Demand features			Statistics	
	Description	AM peak	Trough	PM peak	Cells	Std
Cluster 0	Residential Diffuse	7:05	10:23	14:41	53	0.000919
Cluster 1	Mixed Activity	7:19	9:46	14:16	54	0.000930
Cluster 2	Polycentric Mixed	7:21	9:34	14:57	57	0.000656
Cluster 3	Residential Mixed	7:09	10:02	14:39	107	0.000763

The main difference in cluster profiles are the amplitudes of the peaks, and the relationship between them, showing the method clearly distinguishes key demand features. Closer inspection of the profiles reveals more nuance is also captured, such as details in the onset and tails of the peaks and the inter-peak trough. Note the difference in convexity between the peaks of clusters 2 and 3.
We have labelled *Cluster 0* as *residential diffuse*, it primarily encompasses residential suburban areas primarily serviced bus routes. Its demand profile shows a pronounced morning peak with a shallow and spread-out afternoon peak.
Cluster 1 corresponds to *mixed activity* areas. This cluster correspond mainly to boundaries between dense urban areas and spread-out residential areas. These areas exhibit mixed land use including offices, shopping centres and even part of the Helsinki University campus. The higher afternoon peak is indicative of the demand attraction of the areas, while the convex drop in the afternoon peak shows activities occurring primarily during regular business hours.
Cluster 2, which we have labelled *polycentric mixed*, is the most compact, consisting mainly of the city centre but also extending along the metro line. In terms of its demand profile, it has a significantly higher afternoon peak, indicating it as a demand sink. Looking at scaled demand, however, the morning peak is higher than other clusters' due to the denser development. The decay of the afternoon peak is far more linear than the other clusters, indicative of a wider diversity of trip types.
We have labelled the last cluster (*Cluster 3*) as *residential mixed*. This cluster contains areas close to train stations and, to a lesser degree, the metro stations in eastern Helsinki. These areas exhibit transit oriented development and are densely populated with multi-storey residential buildings and higher concentration of services near PT stops. The higher morning peak reflects the residential aspect, while the mixed nature of the areas — which attracts demand — and proximity to rail stations — where high numbers of boardings occur throughout the whole day — is reflected by the cluster having the most even morning and afternoon peaks.



(a)



(b)

Figure 2: (a) Normalised demand profiles for the different clusters. In lighter grey a sample of demand the profiles of individual cells used for the clustering are shown. (b) Cluster demand profiles scaled by the average number of timestamps per cell.

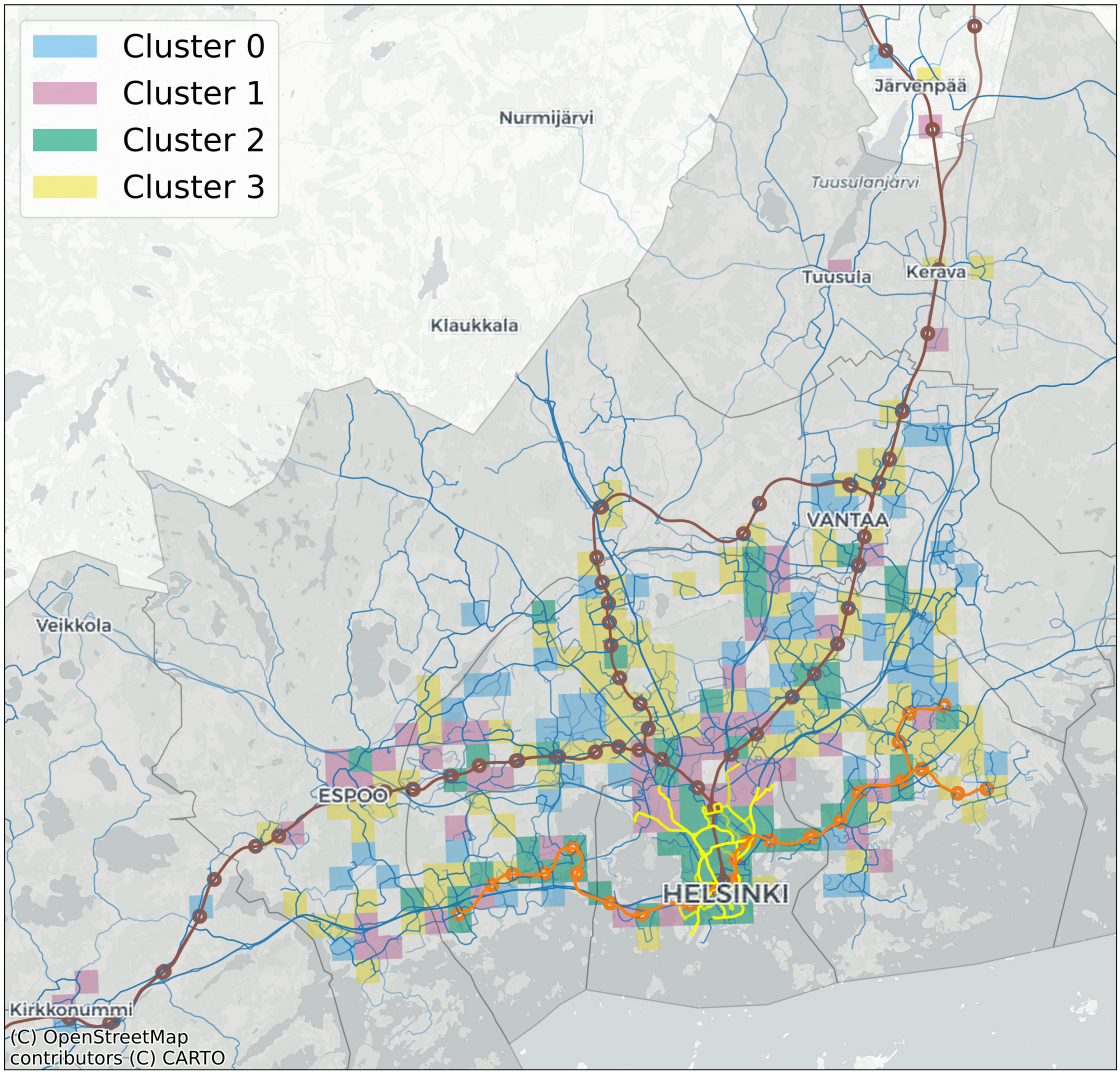


Figure 3: Spatial distribution of cells according to their cluster. The PT lines are shown (buses in blue, train in brown, trams in yellow, and metro in orange) along with the metro and train stations. The grey shaded areas are the fare zones used by HSL which mostly cover the Helsinki metropolitan region.

Supply profiles of demand clusters

Supply profiles for the clusters found based on the demand profiles (see Figure 3). The timestamps for each cluster correspond to the scheduled arrivals of PT vehicles for all the cells in each cluster. Figure 4 shows both the normalised KDE, as well as the scaled supply profiles based on actual volume.

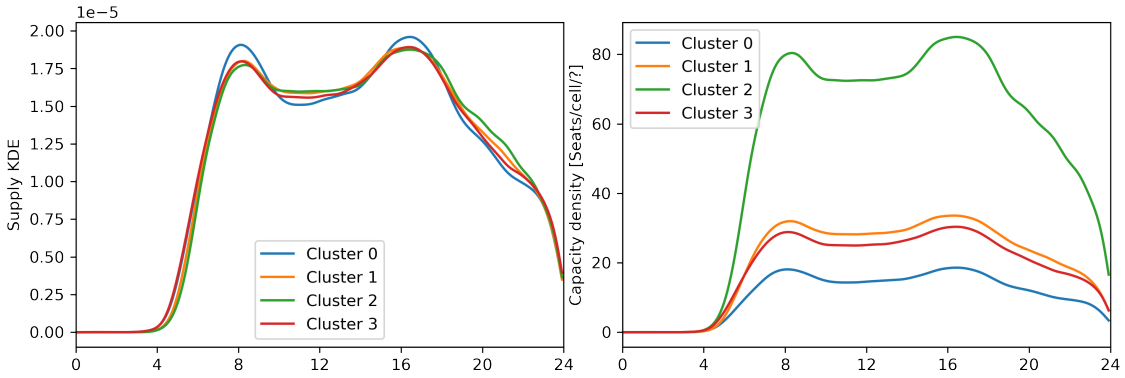


Figure 4: Comparison of the supply profiles for the different clusters. On the left, we the KDE supply profile is shown (normalised to have a unit integral). On the right, the supply profiles are scaled by the average number of capacity per cell in the cluster.

The PT supply is very uniform across the different clusters. The higher supply at peak times is noticeable, however it can be seen that frequency is of services is high during the day at off-peak times as well. This reflect the high frequency of PT services that HSL provides, as the Helsinki metropolitan region has been evaluates as having among the best PT service in Europe.

Discussion

The method presented results in large scale spatial patterns that are derived from temporal transportation patterns at the local level. Furthermore, the clusters coincide with distinct types of urban areas, also revealing the fundamental role of the PT network in shaping demand.
The clustering based only on the shape of the demand profiles (ignoring travel mode and total volumes), allows comparing areas primarily based on the behaviour of the passengers. Our method allows usage of sparse data while preserving profile shape details. The tail-end of the afternoon peaks is revealing of the nature of the urban areas which form each cluster. The variance in the demand patterns also appears to differ significantly for across clusters, opening up additional lines of inquiry.
In terms of supply, the scaled profiles reflect the volumes of the scaled demand curves. Cluster 2 has the highest volumes followed by clusters 1, 3 and 0 respectively. The uniformity across supply profiles indicates that the urban area and travel mode are more important in shaping demand than PT capacity itself.
Limitations of the TravelSense data include undersampling and selection bias Huang et al. (2022). However, a large percentage (43%) of Helsinki residents use PT on a near-daily basis (Brandt et al., 2019), with most passengers using the mobile ticketing app. This could indicate that the issues with the dataset could be easily overcome with increased coverage.

4 Conclusions

We have presented a novel method for reconstructing demand profiles from timestamped event data, and for clustering urban areas based on their PT demand. We demonstrated the use of the method in the Helsinki metropolitan region, with data from the regional transport authority (HSL). The method is very general, which we exemplify by obtaining the supply profiles of transit services using GTFS schedules, highlighting that the method is not restrictive in terms of the data type and formatting required.
In the presented case study, we identify four demand clusters corresponding to qualitatively different types of urban areas, distinctly picking up the dense city centre, suburban, and mixed areas

without relying on passenger volumes and rather on the shape of the daily demand curve for PT journeys.
Our results show that the present method, even with relatively sparse data, yield meaningful spatial patterns derived from the temporal demand profiles. The method allows characterising the demand behaviour of different urban areas with potential application to transport planning.
Our results open up avenues for further research, including higher resolution studies with more data, as well as more rigorous analysis of clustering result's based on the demand profiles.

Acknowledgements

We thank Pekka Räty, Tri Quach and Hanna Kitti from HSL for data access and useful discussions. AEMV was supported by Research Council of Finland project: T-Winning Spaces 2035. ZH was supported by the Strategic Research Council at the Research Council of Finland (grant numbers 345188 and 345183). CS was supported by the FinEst Twins Center of Excellence (H2020 Grant 856602). Calculations were performed using computer resources within the Aalto University School of Science "Science-IT" project.

References

Brandt, E., Kantele, S., & Räty, P. (2019). Travel habits in the helsinki region in 2018. *Helsinki Region Transport. HSL Publications*, *9*, 2019.

Cats, O., & Ferranti, F. (2022). Voting with one's feet: Unraveling urban centers attraction using visiting frequency. *Cities*, *127*, 103773.

Cats, O., Wang, Q., & Zhao, Y. (2015). The identification and classification of urban centres using public transport passenger flows data. *Journal of Transport geography*, *48*, 10–22.

Cuauhtemoc Anda, A. E., & Fourie, P. J. (2017). Transport modelling in the age of big data. *International Journal of Urban Sciences*, *21*(sup1), 19-42.

Devillaine, F., Munizaga, M., & Trépanier, M. (2012). Detection of activities of public transport users by analyzing smart card data. *Transportation Research Record*, *2276*(1), 48–55.

Ewing, R., & Cervero, R. (2010). Travel and the built environment. *Journal of the American Planning Association*, *76*(3), 265-294.

Gonzalez, M. C., Hidalgo, C. A., & Barabasi, A.-L. (2008). Understanding individual human mobility patterns. *nature*, *453*(7196), 779–782.

Hong, J., Shen, Q., & Zhang, L. (2014). How do built-environment factors affect travel behavior? a spatial analysis at different geographic scales. *Transportation*, *41*, 419–440.

Houghton, C. (2015). Calculating mutual information for spike trains and other data with distances but no coordinates. *Royal Society open science*, *2*(5), 140391.

HSL. (2021). *HSL Open data.* `https://www.hsl.fi/en/hsl/open-data#public-transport-network-and-timetables-gtfs`.

HSL. (2021). *Traficom decides to limit public transport passenger numbers – effects on HSL transport services.* Retrieved from `https://www.hsl.fi/en/hsl/news/news/2021/03/traficom-decides-to-limit-public-transport-passenger-numbers-in--effects-on-hsl-transport-services`

Huang, Z., Espinosa Mireles de Villafranca, A., & Sipetas, C. (2022). Sensing multi-modal mobility patterns: A case study of helsinki using bluetooth beacons and a mobile application. In *2022 ieee international conference on big data (big data)* (p. 2007-2016).

Kaufman, L., & Rousseeuw, P. (1990). Partitioning around medoids (program pam). In *Finding groups in data* (p. 68-125). John Wiley & Sons, Ltd.

Lee, S. G., & Hickman, M. (2014). Trip purpose inference using automated fare collection data. *Public Transport*, *6*, 1–20.

Luo, D., Cats, O., & van Lint, H. (2017, January). Constructing Transit Origin–Destination Matrices with Spatial Clustering. *Transportation Research Record*, *2652*(1), 39–49.

Miranda-Moreno, L. F., Bettex, L., Zahabi, S. A. H., Kreider, T. M., & Barla, P. (2011). Simultaneous modeling of endogenous influence of urban form and public transit accessibility on distance traveled. *Transportation Research Record*, *2255*(1), 100-109.

MobilityData IO. (2024). *Open mobility data.* `https://transitfeeds.com/p/helsinki-regional-transport/735?p=56`.

Mouratidis, K., Ettema, D., & Næss, P. (2019). Urban form, travel behavior, and travel satisfaction. *Transportation Research Part A: Policy and Practice*, *129*, 306-320.

Nasri, A., & Zhang, L. (2012). Impact of metropolitan-level built environment on travel behavior. *Transportation Research Record*, *2323*(1), 75-79.

Roth, C., Kang, S. M., Batty, M., & Barthélemy, M. (2011, 01). Structure of Urban Movements: Polycentric Activity and Entangled Hierarchical Flows. *PLOS ONE*, *6*(1), 1-8.

Rousseeuw, P. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, *20*, 53-65.

Sutherland, W. A. (2009). *Introduction to metric and topological spaces.* Oxford University Press.

Thomas, R., & Bertolini, L. (2020). Introduction to transit-oriented development. In *Transit-oriented development: Learning from international case studies* (pp. 1–20). Cham: Springer International Publishing.

Wand, M. P., & Jones, M. C. (1994). *Kernel Smoothing.* CRC press.

Yap, M., Luo, D., Cats, O., van Oort, N., & Hoogendoorn, S. (2019). Where shall we sync? clustering passenger flows to identify urban public transport hubs and their key synchronization priorities. *Transportation Research Part C: Emerging Technologies*, *98*, 433–448.